|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Leeds University**  **Business School** | | A close-up of a sign  Description automatically generated | | | | | | | | |
| **Assessed Coursework Coversheet** | | | | | | | | | | |
| For use with *individual* assessed work | | | | | | | | | | |
| Student ID Number: | 2 | | 0 | 1 | 8 | 0 | 1 | 2 | 9 | 5 |
| Module Code: | LUBS5309M | | | | | | | | | |
| Module Title: | Forecasting and Advanced Business Analytics | | | | | | | | | |
| Module Leader: | Panagiotis Stamolampros | | | | | | | | | |
| Declared Word Count: | 2984 | | | | | | | | | |
| Please Note:  Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such.  If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.  It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices (“appendix abuse”).  It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams. | | | | | | | | | | |
| By submitting an assignment you confirm you have read and understood the University of Leeds [**Declaration of Academic Integrity**](http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf) ( <http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf>). | | | | | | | | | | |

Top of Form

Bottom of Form

**Part 1**

**Introduction**

The aim of this report is to utilise time series modelling and compare the predictive ability of three different forecasting models on the US seasonally-adjusted personal consumption expenditures (PCE) dataset. The goal is to pinpoint the model which is most efficient for precisely forecasting the PCE values.

By comparing and evaluating the performance of these methods, we determine the most suitable approach for time series analysis.

* Evaluating model by assessing the effectiveness of various forecasting methods, including simple methods like average, naïve, seasonal naïve, or drift and more complex models like exponential smoothing and ARIMA, we can assess their capability to identify the underlying trends in PCE data. The filtering method depends on the objectives and characteristics of data as each of these approaches have their own strengths and drawbacks.
* The main objective is to identify the model that generates the most precise forecasts. This entails assessing the effectiveness of each model minimizing accuracy measures.
* The analysis further offers meaningful perspectives for decision-makers in identifying the best model for PCE forecasting.

**Data Description**

The PCE dataset summarizes the spending trends in the United States within a defined time frame. It represents total household and individual spending on goods and services.

1. Given dataset is seasonally adjusted, which means seasonal component is removed from the original dataset to highlight various irregular patterns as well as the underlying trends.
2. PCE observations recorded span from January 1959 to November 2023, with monthly frequency.
3. Dataset contains missing values, which means period having no data value.

**Figure 1:**

**A screenshot of a table

Description automatically generated**

**DATA PRE-PROCESSING**

Time series forecasting is useful in predicting values as it uses historical values and associated patterns to predict future activity. A univariate time series object is created from the given dataset, with start and end dates set from 1959 to 2023 respectively, with monthly frequency.

A check for missing values within the dataset is conducted, to verify data integrity and reliability for precise analysis. To impute the missing values, linear interpolation is applied which helps in maintaining the progression of the time series dataset. It is the simplest type of interpolation, where the missing value is estimated by the mean of the adjacent data values.

**Figure 2:**

A graph with numbers and lines

Description automatically generated

Figure 2, illustrates the time series object generated from the original data.

There is clear upward trend in PCE values indicating gradual increase in consumer spending over the observed period. The inconsistencies from missing values can be observed. Minimal to no seasonality is observed, because data is seasonally adjusted but there are some minor fluctuations.

**Figure 3:**

A graph with a blue line

Description automatically generated

Figure 3 highlights the distribution of missing data in time-series for clearer understanding.

**Figure 4:**

A graph with numbers and a line

Description automatically generated

The figure 4 exhibits the seasonality of the time-series for all the years. This visualization provides clearer insights about the patterns and is useful in identifying years in which the pattern deviates. Similar patterns have been observed for almost all the months throughout the different months, except for APRIL, 2020, where we see a major dip in the PCE value, which could be due to the complex interplay of different factors associated with COVID-19 pandemic like, inflation, containment measures and decline in purchase of non-essential goods, etc.

Another crucial step performed for pre-processing the data before applying the methods was to split the dataset into training and testing section. This is essential for evaluation and validation of the forecasting models. The dataset is distributed into training and testing in an 80-20 split format, where the training set contributes to the model parameters, while the testing dataset assesses the predictive accuracy of the model on unobserved data. Data splitting also helps in preventing biases towards the training data, by ensuring to make more robust evaluations.

In the imputed PCE dataset, we perform this temporal split for training data from January 1959 to December 2010, which falls under 80% of the data, while the remaining from January 2011 to November 2023, were integrated in testing data.

**METHOD ANALYSIS**

Our analysis involves comparison of three models, which are:

1. Naïve Method

It is one of the simplest form of forecasting in which the prediction for the present period is the value of preceding observation of the time series.

When applying naïve method on the training dataset, forecast horizon will be equal to length of testing dataset, as we want to predict values from January 2011 to November 2023 .

**Figure 5:**

A graph showing the growth of a cost

Description automatically generated

In figure 5, The upward trend shows increase in customer spendings over the years. The naïve extends the recent data point into the future horizontally, assuming that the forecast will remain constant at the last observed level. The blue area highlights the uncertainty range, where the actual future values of PCE are likely to fall.

1. Exponential Smoothing

It is a technique that computes the weighted average of historical observations, where weights decrease exponentially on moving further in time. Holt linear or double decomposition method is a simple exponential smoothing technique, that allows time-series forecasting containing trends. The PCE data provided is seasonally adjusted and has upward trend, hence Holt linear method is utilised for analysis.

**Figure 6:**

A graph showing the growth of a number of companies

Description automatically generated

In figure 6, The resulting forecast shows the increasing trend for the testing data potentially at a slightly slower rate than the past. However, relatively wide prediction intervals are seen depicting the variation in the past data.

1. ARIMA

The Autoregressive Integrated Moving Average method is a forecasting approach, which predicts value as a linear combination of its past values and errors in a time series. It makes use of lagged moving averages to smooth time series data.

ARIMA(p,q,d) classification:

p -> number of autoregressive terms

q-> number of lagged forecast errors in equation.

d-> number of non-seasonal differences needed for stationarity.

**Figure 7:**

A graph showing the growth of a cost

Description automatically generated

The forecast in figure 7 shows an upwards trend for the testing data with relatively less prediction intervals depicting that the forecasts are more precise and have low uncertainty.

**METHOD RESULTS**

Before analysing the accuracy metrics, let’s understand another model diagnostic approach, called residual analysis. By analysing the residuals, we check if these residuals have a normal distribution, and if there is autocorrelation between the residuals as well as the residuals to have a zero mean. If these properties are satisfied, the residuals are considered as white noise. The Ljiung-Box test is a statistical test used in time series analysis to validate the presence of autocorrelation in a dataset.

The null hypothesis is that the correlations in the population from which sample is drawn are zero, so that any observed correlations in the data, is an outcome of randomness of sampling process.

**Figure 8:**

**A graph of different types of data

Description automatically generated with medium confidence**A screenshot of a test

Description automatically generated

The above Naïve method, represent the ljung-box test results. The p-value is significantly less than significant value (5%), hence we reject the null hypothesis. This represents that there is autocorrelation present in the model residuals.

**Figure 9:**

**A graph of a graph of a method

Description automatically generated with medium confidence**A screenshot of a computer

Description automatically generated

The p-value for Holt’s linear is 0.06051, which is slightly greater than standard p-value of 0.05. Hence, we cannot reject the null hypothesis, which means, residuals are independent.

**Figure 10:**

**A graph of a graph of a wave

Description automatically generated with medium confidence** A white background with black text

Description automatically generated

The p-value for arima is 21% which is greater than 5%, hence we do not reject null hypothesis.

It is evident, that holt is better than naïve as its p-value is higher. Additionally, arima has the highest p-value, amongst them, but we also check the accuracy metrics such as Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE), Mean Absolute Scaled Error(MASE) to cross validate and check which forecasting model smallest errors.

**Figure 10:**

A screenshot of a computer

Description automatically generated

1. From above figure, we find that the Holt model has smallest errors on the testset as compared to naïve and Arima, and the difference is significant, suggesting that the Holt model performs best across all 3.

**Figure11:**

A graph showing a graph of a number of times

Description automatically generated with medium confidence

Figure 11 demonstrates that Holt(fcholt) appears to be most accurate, followed by the auto-arima(fcauto), while the naive model(fcnaive) performs the worst.

Initially, Arima captured the upward trend reasonably well but then significantly overestimates the actual values towards the end of the forecasting period.

Holt appears to be the most accurate of the three, as it is following the general trend of the actual data closely and doesn't deviate too far from the actual values.

As Naive method assumes that the forecasted values will be same as the last observed value, it remains flat and doesn't capture the upward trend of the actual data.

These observation thus imply, that Holt’s method showcases more accurate predictions due to lower error metrics.

Therefore, Holt’s method becomes the superior model as it enables to capture and generate more precise predictions and patterns.

**OCTOBER 2024 PREDICTIONS:**

As Holt’s linear model demonstrates superior forecasting accuracy by providing lesser errors, we use it to forecast PCE values until October 2024.

**Figure 12:**

A graph showing the growth of a number of times

Description automatically generated

Top of Form

**Bottom of Form**

**ROLLING ORIGIN**

To cross validate the best model, we use rolling origin approach. It is an evaluation technique, wherein the forecasting origin is continuously updated and for each new origin, the forecast is generated. This method gathers several forecasting errors for time series, thus providing more robust performance estimate.

**Figure 13:**

A screenshot of a computer

Description automatically generated

The output compares three models, after applying one step rolling:

* While naïve has largest errors and Arima incorporated moderate error values, Holt, showed the lowest values across all metrics.
* **Holt Linear** significantly outperforms both the Naive and ARIMA models based on the accuracy metrics. This suggests that Holt Linear effectively captures the linear trend in the data.
* The Naive and ARIMA models have identical performance, which indicates that ARIMA did not find any suitable parameters for this specific dataset.

**Conclusion:**

This output provides valuable insight into the performance of different time series forecasting models. It highlights the importance of model selection and the benefits of using rolling origin validation for robust evaluation.

**Limitations:**

1. The availability of old data, can cause uncertainty and may reduce the effectiveness of forecasting models.
2. The auto.arima() only explores the predefined model sets and do not search throughout for all possible combinations to recognise best model. Hence, there might be some unexplored models, which could provide better results.

Part 2:

Introduction

The primary objective of this report is to analyse online hotel reviews and perform topic modelling to explore the customer feedback. These reviews, with corresponding ratings, provides valuable insights into customers' experiences. In this report, we will work on a small sample of data, wherein we split it into negative and positive review based on the ratings. We will further perform topic modelling, to product the top 3 factors for both negative and positive feedback.

Data Description

Our dataset comprises 10,000 rows, containing customer reviews for the hotel highlighting their experience, along with a Likert scale rating ranging from 1 (low satisfaction) to 5 (high satisfaction).

The reviews are multi-lingual, depicting diverse customer perspective, incorporating both positive and negative sentiments.

Data pre-processing

Using the whole dataset, we detect the reviews that are in ‘English’ only, to get better understanding of the data. Out of the 7,982, English reviews we, create a random sample set of 2,000 reviews from these reviews using the sample\_n() function , to ensure the robustness of our analysis. This process of filtering out non-English reviews helps in managing dataset in a consistent language, allowing more efficient and accurate analysis. Furthermore, we prepared the sample data by filtering the punctuations and making sure that there's space between numeric values and alphabetic characters as, separating them ensures that numbers are considered as separate tokens, which will further help in better analysis. To analyse the data further, different analysis techniques have been used.

1. **Sentiment Analysis**

Sentiment analysis is the process of analysing qualitative data to determine if the emotional tone of the text is positive, negative, or neutral.

Using this analysis, the sentiment score gets calculated, by which we split the reviews. The scores typically range from -1 to 1, therefore

* If sentiment score > 0 , positive review.
* If sentiment score <= 0, negative review

This pre-processing step helped in analysing the sentiment of the reviews and understanding their overall sentiment distribution.

**Figure 14:**

A graph of a normal distribution

Description automatically generated

The above visualisation shows, that the sample data contains more positive reviews than negative reviews. The mean sentiment value of 0.6584876 is detected, which also indicates that the overall sentiment of the sample text data is positive, as the average sentiment score is closer to 1, thus suggesting that customers have positive opinions about the hotels being reviewed.

**Figure 15:**

A graph of a positive and negative result

Description automatically generated with medium confidence

Top of Form

Bottom of Form

Out of the 2,000 sample data, 1817 values show positive reviews, while 183 values show negative reviews.

1. **Text Pre-processing**

In text analysis, we need to convert the text documents into a corpus to allow structured way to represent text data. It organizes texts in a consistent manner, which helps in easier pre-processing and analysis. Each document in the corpus is considered as a separate analysis unit, enabling efficient processing and analysis of large collections of documents.

Two corpus for the positive and negative reviews are made, in order to conduct text analysis techniques like topic modelling. Furthermore, with the two corpus created, we performed data cleaning, by conducting lemmatization, removal of punctuation, numbers, stopwords, and converting tokens in lowercase.

To perform these transformations, we use tm\_map() function, which helps in mapping a function to all the elements of the corpus.

* Lemmatization is term transformations that help grouping similar tokens together. Using this we reduce the vocabulary size which helps in improving the text analysis.
* Stopwords removal focuses on the content-carrying words which provide more informative for analysis, while reducing noise. Removing stopwords, helps in improving the efficiency and effectiveness of the text analysis by focusing on the main information.

1. **Document Term Matrix (DTM)**

A DTM matrix is created for both positive and negative corpus after pre-processing is completed to describe frequency of terms in a collection of documents. The rows corresponds to documents columns corresponds for terms in the document.

The value of the matrix depicts the frequency of occurrence of a particular term in a specific document.

**Figure 16:**

A white background with black numbers

Description automatically generated

The above output shows, that there are 1,817 rows of positive reviews, which contains 5,904 tokens(words) in total with sparsity of 99%. While the negative dtm shows, that there are 183 rows, containing 1,907 terms in total with sparsity of 98%.

A new frequency matrix is created that returns information about the repeated words in the dtm, that is further utilised in visualization via wordcloud and LDA at the end.

1. **Visualisation**

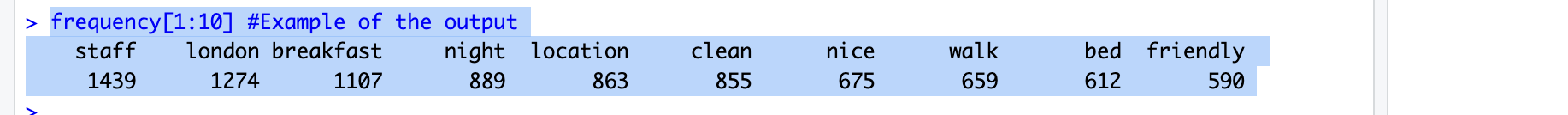
A wordcloud is a visual representation, that takes tokens as input and the associated frequencies. The top ten words, with their corresponding frequencies have been shown below.

**Positive Reviews**

**Figure 17:**

A close up of words

Description automatically generated

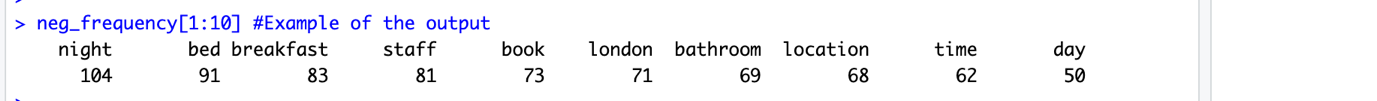


**Negative Reviews**

**Figure 18:**

A close up of words

Description automatically generated



**Methodology**

Topic Modelling is an unsupervised machine learning approach that can analyse document collection, find tokens and phrase patterns within them, and automatically groupings words and related expressions into topics. Latent Dirichlet allocation(LDA) is the one of the most common algorithms for topic modelling, and follows two principles:

* Each document has a mixture of topics.
* Each topic is a mixture of words.

LDA estimate both of these principles simultaneously and find the combination of words that associated with each topic.

After the corpus and pre-processing of the terms, the next step is the selection of number of topics(k), that LDA creates using the ldatuning package. Additionally, we want to minimize Arun2010 and CaoJaun2009 criteria and maximize the Griffiths2004, to determine the number of topics

**Figure 19:**

A graph of a number of topics

Description automatically generated

From the above plot, we can understand that, for positive dtm, the minimum values for Arun2010 and CaoJaun2009 is at 14, and the Griffiths2004 is the maximum. Therefore, we need k=14 topics for analysis.

**Figure 20:**

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

The above plot for negative dtm showcases number of topics at 13 depicting minimum Arun2010 and CaoJaun2009 and Griffiths2004 having maximum value. Therefore, we need k = 13 for LDA.

**Results and discussion**

Using the values of k, we get the below topics:

1. **Positive Reviews**

**Figure 21:**

A screenshot of a computer

Description automatically generated

**Figure 22:**

**A screenshot of a graph

Description automatically generated**

The figure, illustrates the positive topics created by topic visualisation. The bubbles represent the different topics, and the distance between them represents how similar or different the topics are. Closer the circle, more similar the topic.

Based on this, the top three factors impacting customer satisfaction are:

1. **Topic 4 Staff** **Management**: Frequency words like “staff,” “friendly,” “helpful,” and “efficient” indicates that the behaviour and attitude of staff is critical in shaping customer experiences. Customers value friendly, welcoming, and helpful service. Quick request handling techniques also contributes to a positive experience.
2. **Topic 7 Location & Surroundings:** Words like “walks,” “streets,” “shops,” “restaurants,” and “parks” indicates the positive aspects of hotel’s surroundings and ease of exploring it.
3. **Topic 13 Accommodation quality:** Positive customer sentiments associated with hotel amenities are revealed. Customers enjoy stunning views, convenient access, and spacious accommodations. Upgrades to suites are appreciated, along with well-built facilities and inviting lounge areas, thereby enhancing the overall experience during their stay.

**Top of Form**

**Bottom of Form**

1. **Negative Topics**

**Figure23:**

A screenshot of a computer program

Description automatically generated

**Figure 24:**

A screenshot of a graph

Description automatically generated

Similar to figure 22, the above graph shows the topic visualisation for negative topics.

The topics that caused negative impact on the customer satisfaction are :

1. **Topic 1 Room Experience:** Frequency words like “noise,” “noisy,” “uncomfortable,” and “hot” indicates issues related to noise disturbances, uncomfortable sleep quality, poor air quality, and poor room, particularly in relation to the window, bathroom, and sleep quality.
2. **Topic 2** **Check-in & Stay Experience:** improper check-in, cramped spaces, overcrowding, washrooms conditions and blockage, disturbances during night, internet issues, and unfavourable location suggesting customer's initial impressions during their stay.
3. **Topic 3** **Staff Behaviour & Reception**: words like “rude”, “receptionist”, “desk”, “receptionist” highlights encounters with receptionists, observations of staff behaviour, and image of the central desk, illustrating the impact of personnel interactions on guest satisfaction and overall experience.

Top of Form

Bottom of Form

**RESULTS**

Overall, factors such as friendly staff interactions, accessible hotel locations, and high-quality accommodations positively impact customer satisfaction, leading to memorable and enjoyable experiences for guests. While factors like, room quality, check-in experiences, and staff behaviours towards guests have a detrimental effect on customer satisfaction, resulting in dissatisfied guests which potentially affect the hotel's reputation and business success.

References –

1. Investopedia. (n.d.). *Understanding Time Series*. [online] Available at: <https://www.investopedia.com/terms/t/timeseries.asp#:~:text=In%20particular%2C%20a%20time%20series>.
2. ibf.org. (n.d.). *Naïve Model | IBF*. [online] Available at: <https://ibf.org/knowledge/glossary/naive-model-197>.
3. courses.worldcampus.psu.edu. (n.d.). *Naïve Approach*. [online] Available at: <https://courses.worldcampus.psu.edu/welcome/mangt515/lesson02_08.html>.
4. Investopedia. (n.d.). *Autoregressive Integrated Moving Average (ARIMA)*. [online] Available at: <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=Autoregressive%20integrated%20moving%20average%20(ARIMA)%20models%20predict%20future%20values%20based>.
5. Zach (2022). *How (And When) to Use set.seed in R*. [online] Statology. Available at: <https://www.statology.org/set-seed-in-r/>.
6. AWS (2023). *What is Sentiment Analysis? - Sentiment Analysis Explained - AWS*. [online] Amazon Web Services, Inc. Available at: <https://aws.amazon.com/what-is/sentiment-analysis/>.
7. Tian, Y. (n.d.). *3 Text Analytics with R: Topic Modelling | AI4PH: Text Analyses with R Tutorial and Data Challenge*. [online] *bookdown.org*. Available at: https://bookdown.org/tianyuan09/ai4ph2022/tutorial2.html#topic-modelling [Accessed 12 May 2024].
8. R-Packages. (n.d.). *How can this be done?* [online] Available at: <https://cran.r-project.org/web/packages/greybox/vignettes/ro.html>.
9. rpubs.com. (n.d.). *RPubs - Time Series III*. [online] Available at: https://rpubs.com/Pstamo/Tseries3New [Accessed 12 May 2024].
10. 7.1 Simple exponential smoothing | Forecasting: Principles and Practice. (n.d.). [online] *otexts.com*. Available at: <https://otexts.com/fpp2/ses.html>.
11. www.mathworks.com. (n.d.). *Ljung-Box Q-test for residual autocorrelation - MATLAB lbqtest*. [online] Available at: https://www.mathworks.com/help/econ/lbqtest.html [Accessed 12 May 2024].